**NYU Data Science Community** features journalism, research papers, events, tools/software, and jobs for May 27, 2016

Please let us (Laura Noren, Brad Stenger) know if you have something to add to next week's newsletter. We are grateful for the generous financial support from the Moore-Sloan Data Science Environment and to NYU's Center for Data Science.

## Data Science News

### NASA Helps Launch Data Science Grad Program

*datanami* from May 23, 2016
Another U.S. university is adding a data science specialization to its curriculum, this one as part of an online Masters of Science degree in engineering. The **University of California at Riverside** said the data science track was developed in collaboration with NASA's **Jet Propulsion Laboratory** science staff.

Also, in university data science:

- Artificial Intelligence Authority Named Jacobs Technion-Cornell Institute Director (May 25, The Cornell Daily Sun )
- Why I Chose Cornell Tech and Jacobs (May 25, Cornell Tech, News & Views; Ron Brachman)

### Astronomers ink deal to build record telescope

*Science, ScienceInsider* from May 25, 2016
Astronomers today signed an unprecedented contract to build the world's largest ground-based optical and infrared telescope. In a ceremony at the headquarters of the **European Southern Observatory** (ESO) in Garching, Germany, ESO Director General Tim de Zeeuw inked the record deal—worth €400 million—with three Italian engineering firms. They will build the structure that will hold the huge 39-meter mirror of the European Extremely Large Telescope (E-ELT), as well as the domed building that will enclose it.

### Visualization of Publication Impact

*arXiv, Computer Science > Digital Libraries; Eamonn Maguire, Javier Martin Montull, Gilles Louppe* from May 20, 2016
Measuring scholarly impact has been a topic of much interest in recent years. While many use the citation count as a primary indicator of a publications impact, the quality and impact of those citations will vary. Additionally, it is often difficult to see where a paper sits among other papers in the same research area. Questions we wished to answer through this visualization were: is a publication cited less than publications in the field?; is a publication cited by high or low impact publications?; and can we visually compare the impact of publications across a result set?

### HIPAA doesn't apply to Precision Medicine Initiative, sparking privacy concerns

*Becker's Health IT and CIO Review* from May 20, 2016
A central concern to genomics is the aggregation of personal health information in one place. A report from the **World Privacy Forum** expounds upon this concern, and

others, and suggests the federal government's Precision Medicine Initiative is too ambiguous and lax on its privacy guidelines.

Chief among the concerns is that medical record data and biospecimen data contributed to the initiative are not covered under HIPAA.

Also, in Precision Medicine:

- White House releases final Precision Medicine Initiative data security framework (May 26, Healthcare IT News)

### Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.

*ProPublica* from May 23, 2016

... something odd happened when Borden and Prater were booked into jail: A computer program spat out a score predicting the likelihood of each committing a future crime. Borden — who is black — was rated a high risk. Prater — who is white — was rated a low risk.

Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics.

### Why the Very Silly Oracle v. Google Trial Actually Matters

*VICE, Motherboard* from May 25, 2016

The jury is currently in deliberations over whether Android's use of the declaring code, and the structure, sequence, and organization of 37 Java API packages, was fair use.

The precedent that APIs are copyrightable was already set by the Federal Circuit in 2014. But the US Copyright Act doesn't protect purely functional things. It doesn't extend to any "process, system, method of operation." If it seems to you like APIs are so purely functional that they should be covered under that latter doctrine, you're not alone.

Also, in Oracle vs. Google:

- Google beats Oracle—Android makes "fair use" of Java APIs (May 26, Ars Technica, Joe Mullin)

### Facebook and Microsoft Are Laying a Giant Cable Across the Atlantic

*WIRED, Business* from May 26, 2016

**Facebook** and **Microsoft** are laying a massive cable across the middle of the Atlantic.

Dubbed MAREA—Spanish for "tide"—this giant underwater cable will stretch from Virginia to Bilbao, Spain, shuttling digital data across 6,600 kilometers of ocean. Providing up to 160 terabits per second of bandwidth—about 16 million times the bandwidth of your home Internet connection—it will allow the two tech titans to more efficiently move enormous amounts of information between the many computer data

centers and network hubs that underpin their popular online services.

### What's Next for Digital Humanities?

*Communications of the ACM* from June 01, 2016
Today, digital humanists are applying advanced computational tools to a wide range of disciplines, including literature, history, and urban studies. They are learning programming languages, generating dynamic three-dimensional (3D) re-creations of historic city spaces, developing new academic publishing platforms, and producing scholarship.

The breadth of the field has led to something of an identity crisis.

Also, in digital humanities:
- The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics (May 23, CA: Journal of Cultural Analytics; Lev Manovich)
- What Digital Humanists Do (May 25, Software Carpentry)

### Deep biomarkers of human aging: Application of deep neural networks to biomarker development

*AGING Journal; Alex Zhavoronkov et al.* from May 18, 2016
One of the major impediments in human aging research is the absence of a comprehensive and actionable set of biomarkers that may be targeted and measured to track the effectiveness of therapeutic interventions. In this study, we designed a modular ensemble of 21 deep neural networks (DNNs) of varying depth, structure and optimization to predict human chronological age using a basic blood test. To train the DNNs, we used over 60,000 samples from common blood biochemistry and cell count tests from routine health exams performed by a single laboratory and linked to chronological age and sex.

### Attention Kid Scientists! – The President Wants Your Ideas on Science and Technology

*The White House* from May 19, 2016
**President Obama** wants to hear from YOU – kid scientists and innovators across the country – about what we can do to help shape the future of science, discovery, and exploration.

Also, from the White House:
- Administration Issues Strategic Plan for Big Data Research and Development (May 23, The White House, Keith Marzullo)

### When it Comes to Replicating Studies, Context Matters, an Analysis of Reproducibility Project Work Finds

*NYU News* from May 23, 2016
Contextual factors, such as the race of participants in an experiment or the geography of where the experiment was run, can reduce the likelihood of replicating psychological studies, a team of **New York University** researchers has found. Their work, which appears in the journal *Proceedings of the National Academy of Sciences*,

analyzed papers examined by the Reproducibility Project in an effort to identify potential challenges to replicating scientific scholarship.

Also, in context and reproducibility:

- Context matters when replicating experiments, argues study (May 23, Retraction Watch)
- Contextual sensitivity in scientific reproducibility (May 23, Proceedings of the National Academy of Sciences; Jay J. Van Bavel, Peter Mende-Siedlecki, William J. Brady, and Diego A. Reinero)

### Tweet of the Week: What idiot called it "deep learning hype" and not "backpropaganda"

*Twitter, Naomi Saphra* from April 14, 2016




## Events

### Artificial Intelligence For Social Good workshop

Organizers: **The White House** and the **Computing Community Consortium**

**Washington, DC** Tuesday, June 7, at The Willard Intercontinental Hotel

### CUSP Research Seminar Series | June 15

Join **NYU Center for Urban Science and Progress** for a research seminar with **Jeff Jonas**, **IBM** Fellow and Chief Scientist of Context Computing.

**Brooklyn, NY** on Wednesday, June 15, at 11 a.m., Jacobs Seminar Room at the Center for Urban Science and Progress (1 Metrotech, 19th Floor)

### Workshop: The Science of Data-Driven Storytelling

Organizers: The National Science Foundation's **West Big Data Innovation Hub** and **DataScience, Inc.**

**Culver City, CA** Thursday, June 16, from 12-4 p.m., at 200 Corporate Pointe, Suite 200 in Culver City

### Accepted Papers - Data-Efficient Machine Learning workshop at ICML 2016

This ICML 2016 workshop will discuss the diversity of approaches that exist for data-efficient machine learning, and the practical challenges that we face.

**New York, NY** Friday, June 24, at the Marriott Marquis in Times Square [$]

### Talk & Poster List | SciPy 2016

The following Talks and Posters have been accepted for SciPy 2016. Additional Talks and Posters will be announced soon.

**Austin, TX** Monday-Sunday, July 11-17. [$]

## Deadlines

---

### Open Data Research Symposium 2016 Call for Abstracts

The Open Data Research Symposium program committee is pleased to announce that the second Open Data Research Symposium (#ODRS16).

**Madrid, Spain** Wednesday, October 5, will be held on October 5, 2016 prior to the International Open Data Conference 2016.

The deadline for abstracts submissions is Monday, May 30.

### O'Reilly Artificial Intelligence Conference, September 26 - 27

The O'Reilly Artificial Intelligence Conference Call for Speakers is open. Also, see the essay, Why AI is finally going mainstream, by **Tim O'Reilly**.

**New York, NY** on Monday-Tuesday, September 26-27.

Deadline to apply to speak is Monday, June 6.

### Text as Data 2016 | Seventh Annual New Directions in Analyzing Text as Data

We invite you to submit a paper to be presented, or to submit your name as discussant or as an attendee, at the seventh annual research conference on "New

Directions in Analyzing Text as Data" that will be held at **Northeastern University** on October 14-15, 2016.

**Boston, MA** Friday-Saturday, October 14-15 at Northeastern University.

Deadline for abstract submissions is Monday, June 20.

### George B. Dantzig Dissertation Award

The George B. Dantzig Award is given by **INFORMS** for the best dissertation in any area of operations research and the management sciences that is innovative and relevant to practice. This award has been established to encourage academic research that combines theory and practice and stimulates greater interaction between doctoral students (and their advisors) and the world of practice.

Deadline to submit 2016 applications is Thursday, June 30.

### NEW: The International Prize in Statistics

A joint effort created by five international statistical societies—**American Statistical Association**, **International Biometric Society**, **International Statistical Institute**, **Institute of Mathematical Statistics**, and the **Royal Statistical Society**—the International Prize in Statistics also raises awareness of the invaluable role that statistics, data analysis, probability, and the understanding of uncertainty play in science, technology, human welfare, and the overall advancement of society.

Deadline for submissions is Monday, August 15.

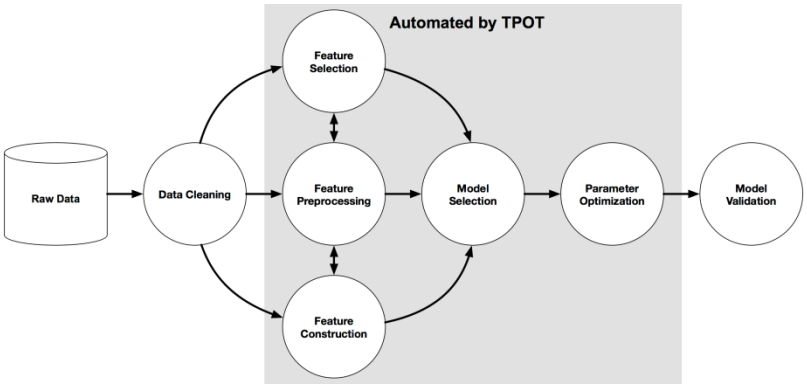
## Tools & Resources

### bandicoot, an open-source python toolbox to analyze mobile phone metadata

*MIT Media Lab* from May 06, 2016
We released a new version (0.5) which includes an interactive visualization, support for mobile phone recharges, support for Python 3, and clustering algorithms to handle both antenna and GPS locations. The computations are significantly faster and the memory footprint is reduced. This release is available on GitHub and PyPI.

### TPOT

*GitHub - rhiever* from February 23, 2016
Consider TPOT your Data Science Assistant. TPOT is a Python tool that automatically creates and optimizes machine learning pipelines using genetic programming.




## How predictive APIs are used at Upwork, Microsoft and BigML (and how they could be standardized)

*Papis.io, PAPIs stories* from May 23, 2016
Predictive Application Programming Interfaces (APIs) are receiving a lot of interest in the industry as they accelerate the development of predictive applications, by making it easier for developers to use predictive models in production settings. They are a means of exposing predictive models to other programs, and they can also expose model learning capability, in which case one may speak of Machine Learning (ML) APIs. They exist in commercial offerings, such as Microsoft Azure ML, BigML, Amazon ML, Datagami and Google Prediction API, where ML algorithms run on cloud platforms and are accessed “as a service” (MLaaS). Predictive APIs can also be created from open-source or custom frameworks and be self-hosted, as presented by Upwork and PSI (see below), as well as Seldon and PredictionIO (recently acquired by Salesforce).

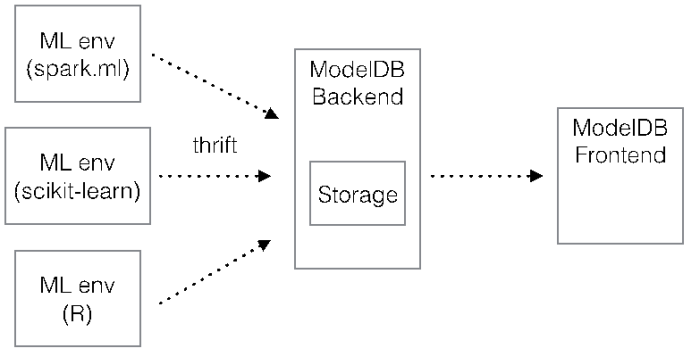
## Deploying Elasticsearch at Scale for Social Media Analytics

*Spinn3r blog* from May 23, 2016
Earlier today we launched a major new release of Spinn3r. This has been in development for about a year so it's really great to get it over the fence and released and in front of customers.

I finally wanted to take some time and write up some details of our Elasticsearch infrastructure which I think would be interesting to other startups and companies in the space.

## ModelDB: A System for Managing Machine Learning Models

*Intel Science & Technology Center for Big Data* from May 25, 2016
We are building a novel end-to-end system called ModelDB to manage ML models. ModelDB clients automatically track machine learning models in their native environments (e.g., scikit-learn, spark.ml), the ModelDB backend introduces a common layer of abstractions to represent models and pipelines, and the ModelDB frontend allows visual exploration and analyses via a web-based interface. Figure 1 shows the high-level architecture of ModelDB.




### Open Sourcing Twitter Heron

*Twitter Blogs, Karthik Ramasamy* from May 25, 2016
Last year we announced the introduction of our new distributed stream computation system, Heron. Today we are excited to announce that we are open sourcing Heron under the permissive Apache v2.0 license. Heron is a proven, production-ready, real-time stream processing engine, which has been powering all of **Twitter's** real-time analytics for over two years. Prior to Heron, we used Apache Storm, which we open sourced in 2011. Heron features a wide array of architectural improvements and is backward compatible with the Storm ecosystem for seamless adoption.

## Careers

### Postdoctoral position in Machine teaching and human learning at Rutgers University -- Newark, Cognitive and Data Sciences Lab

*Rutgers University -- Newark* from May 13, 2016
Newark, NJ. Applications are now being accepted for a post-doctoral fellowship in Dr. **Patrick Shafto's** lab investigating computational models for automating teaching.

The position is ideal for a someone with a strong technical background such as Ph.D. in computer science, cognitive science, mathematics or physics. Strong experimental skills would also be a plus.

If interested, email Dr. Shafto, patrick.shafto@gmail.com, including a CV.

### Researcher (EBM Data Lab) at University of Oxford

*University of Oxford* from May 23, 2016
**Oxford, England** The EBM Data Lab at the University of Oxford is a new project creating innovative, live, impactful projects using academic and health data, led by Dr **Ben Goldacre**. Alongside publishing academic papers, we are also building a range of live data tools with direct clinical and scientific impact such as www.OpenPrescribing.net.

We are seeking a researcher to contribute to a range of projects, collaborate on writing papers, help design and run analyses, and develop ideas into grant proposals. There will also be opportunities to teach and supervise students. You will be working

on a range of our projects on topics including prescribing data, outlier detection, publication bias, research integrity, retractions, healthcare dashboards, and more

### Post-doctoral Position in Bioinformatics at Yale

*Gerstein Lab, Yale University* from May 23, 2016
**New Haven, CT** Applicants are invited for a post-doctoral position at **Yale University**. The position is for two years with possible extensions. Choice of project would to some degree depend on the applicant's interests and abilities, though it is expected that the research will be purely computational and will fall into sub-areas in computational biology and bioinformatics such as biological networks, personal genomics, analysis of next-generation sequencing, cancer genomics and macromolecular structure. The technical approach taken will very much emphasize data mining, machine learning and data science. For specifics, see lab research areas. A step-by-step guide to learning about the lab for prospective postdocs is available.

### Postdoc positions in Social Data Science at Centre for Social Data Science

*University of Copenhagen* from May 18, 2016
Applications are invited for two 2-year postdoc positions at the Faculty of Social Sciences, **University of Copenhagen**. Employment is scheduled to begin on July 1, 2016 or as soon as possible.

The Centre for Social Data Science (SODAS) is a new interdisciplinary center at the Faculty of Social Sciences, University of Copenhagen, involving faculty and students from anthropology, economics, political science, psychology and sociology, as well as from DTU Compute. At the heart of the Faculty's Social Big Data Initiative, SODAS' vision is to create a social data science community focused on leveraging advances in data science and in the collection of digital and/or big data and new data forms for the benefit of social scientists and, at the same time, to study how such data, and not least their ethical and privacy-related challenges, transform the way of doing social science.

### Insight Health Data Science Fellows Program Expands to Silicon Valley

*Insight Data Science* from May 26, 2016
**San Francisco, CA** The Insight Health Data Science Fellows Program is expanding to Silicon Valley this fall to help PhDs and MDs enter the field of health data science. Building on the success of the Boston-based Health Data program, we are excited to help Insight Fellows on the west coast enter this emerging specialization of data science in the life sciences, medicine, and healthcare.